# LAPS - Localisation using Appearance of Prior Structure: 6-DoF Monocular Camera Localisation using Prior Pointclouds

Alexander D. Stewart, Paul Newman

***Abstract*— This paper is about pose estimation using monocular cameras with a 3D laser pointcloud as a workspace prior. We have in mind autonomous transport systems in which low cost vehicles equipped with monocular cameras are furnished with preprocessed 3D lidar workspaces surveys. Our inherently cross-modal approach offers robustness to changes in scene lighting and is computationally cheap. At the heart of our approach lies inference of camera motion by minimisation of the Normalised Information Distance (NID) between the appearance of 3D lidar data reprojected into overlapping images. Results are presented which demonstrate the applicability of this approach to the localisation of a camera against a lidar pointcloud using data gathered from a road vehicle.**

## I. Introduction

This paper is motivated by the expectation that any practical system for future autonomous passenger vehicle navigation will make extensive use of prior information gathered from multiple sources using multiple modalities; such as that available from Google Street View[1] [1]. We further assume that this prior information will be available during live operation, that its coverage will increase to cover the vast majority of areas in which passenger vehicles will operate and that it can be pre-processed offline to improve its utility for the task at hand (navigation in this case).

We argue that these assumptions are both feasible and reasonable, given the already prevalent coverage of services such as Google Street View, mobile data networks and the rapid decline in cost and increase in capabilities of cloud computing and data storage. Mass-production may reduce the cost of large, multi-modality sensor arrays such as those used in the DARPA Urban Challenge [2], [3], to the point where they would be cost-effective for passenger vehicles. However, the model used by Google Street View, whereby a small number of survey vehicles are equipped with expensive sensor arrays and then used extensively to map the environment seems significantly more efficient, as it should result in reduced sensor requirements and thus costs, for subsequent vehicles.

Under these assumptions, we seek a new localisation system suitable for long-term autonomy of passenger vehicles with the following requirements:

R1. *Real-time, robust & accurate estimation of 6-DoF pose relative to the prior information reference frame.*

R2. *Use of incrementally updatable, pre-processed prior information suitable for online use.*

Authors are from the Mobile Robotics Group, Department of Engineering Science, Oxford University {alexs,pnewman}@robots.ox.ac.uk

[1] http://maps.google.com/help/maps/streetview

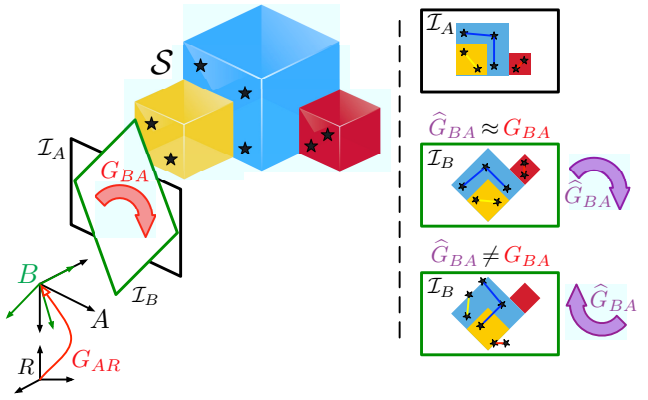



Fig. 1. Two-view reprojection example, the camera centers are denoted by the coordinate frames $\{A, B\}$, the images by $\{\mathcal{I}_A, \mathcal{I}_B\}$; $G_{BA} \in SE(3)$ is the homogenous transform that takes points in frame $A$ and maps them to points in frame $B$: $\boldsymbol{q}_B \equiv G_{BA} \cdot \boldsymbol{q}_A$. Assuming $\widehat{G}_{AR}$ is known, when the estimated $\widehat{G}_{BA}$ is accurate ($\widehat{G}_{BA} \approx G_{BA}$) the reprojection into $\mathcal{I}_B$ of the points in $\mathcal{S}$ align with the captured image and thus their appearance is consistent in both images. When the estimate is wrong ($\widehat{G}_{BA} \neq G_{BA}$), the resulting reprojected points do not align with $\mathcal{I}_B$ and their appearance is inconsistent.

R3. *Robustness to large changes in environmental conditions (day/night/rain etc).*

R4. *Use of low-cost sensors.*

In this paper we formulate the localisation problem as an optimisation over the robot's pose to harmonise the locally observed appearance of known 3D structure gathered previously. Intuitively we want the appearance of known 3D points in the world as viewed from the robot to be localised, to be approximately constant as the robot moves past them as shown in Fig. 1. We assume that the environment has previously been extensively mapped by a separate survey vehicle equipped with a 3D lidar [4], high-resolution cameras and a high-quality INS (Inertial Navigation System). The robot to be localised is assumed to be operating wholly within the mapped area and to be equipped with a collection of low-cost cameras. We make no assumptions about the external conditions during the robot's traversal relative to those of the survey vehicle's; nor do we make any assumptions about the specific sensor configuration used by the robot relative to that used by the survey vehicle(s).

## II. Related Work

In the context of SLAM, 'direct' approaches such as [5] have been proposed that minimise a function of the difference in pixel intensities between a warped reference image patch

and the currently observed image. In [5], multiple planar patches were tracked and their corresponding homographies estimated by minimisation of the $L_2$ norm of pixel intensity differences.

Recent work [6] utilises a direct approach to the problem of visual odometry using a stereo camera, the disparities from which are used via quadrifocal geometry to warp the reference images from the stereo camera prior to comparison with the current images. This approach allows for the accurate and tractable warping of complex scene geometries not feasible by planar homographies. The cost function minimised is again the $L_2$ norm, but with a robust Huber kernel. Our problem differs in that we want to perform localisation against a prior map using only monocular cameras with no locally-sensed range information, rather than visual-odometry using only locally sensed range information.

Perhaps the work for which the guiding intuition is most similar to ours is the localisation component of DTAM (Direct Tracking and Mapping) [7], in which the image feed is used in parallel to both estimate the 3D structure of the local scene and localise the camera within the scene. The latter is achieved by minimising the $L_2$ norm of the photometric error between the current image and a synthesised view of the estimated 3D environment model from a hypothesised camera position. However, currently no external prior information is used in DTAM and the type and scale of the environment considered: indoors and near-field; is very different to ours, outdoor localisation of a moving vehicle at highway speeds over large temporal (seasonal) and physical (city) scales.

Direct image registration by minimisation of the mutual information of the pixel intensities is a widely utilised technique in the context of medical imaging [8] and is known to exhibit significant robustness to both illumination and modality changes [9] in addition to being well suited to minimisation via numerical techniques [10]. Such techniques have been used successfully for 3D object tracking with a monocular camera using a prior model [11], a dual of the problem of robot localisation against a prior map we consider here. Recently they have also been applied successfully to the problem of visual servoing [12], [13] the latter in the context of a ground vehicle tracking a previous route defined by a sequence of previously captured images.

## III. Localisation using Appearance of Prior Structure

### A. Problem Formulation

Consider the case of a robot with a camera moving through an arbitrary, known 3D scene $\mathcal{S}$ taking images at different positions, as shown in Fig. 2. Considering the left-hand side, we wish to estimate the current position of the robot: $B$, with respect to the reference coordinate system for the local scene $R$ through the composition of the previous position $A$, in $R$: $G_{AR}$ and the motion between the previous and current frames: $G_{BA}$.

The intuitive hypothesis for our approach is that the true values of $G_{AR}$ and $G_{BA}$ are those which harmonise the

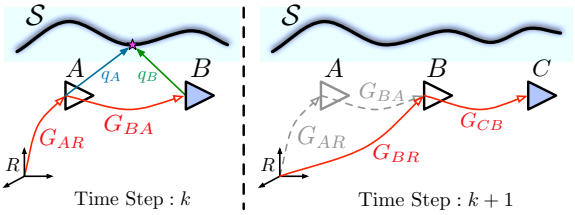



Fig. 2. A robot equipped with a camera observes the same scene $\mathcal{S}$, defined in reference coordinate system $R$ at positions $\{A, B, C, \dots\}$ at which it captures images $\{\mathcal{I}_A, \mathcal{I}_B, \mathcal{I}_C, \dots\}$. For each new image acquired, the transforms to be estimated are shown in red and the current pose in blue.

information about the appearance of $\mathcal{S}$ provided by $\mathcal{I}_A$ & $\mathcal{I}_B$, which describe the appearance of overlapping subsets of $\mathcal{S}$ in some local neighbourhood. Informally, we seek the $G_{AR}$ & $G_{BA}$ which minimise the difference in appearance of known points in the world viewed from different, local, view-points.

In the context of a ground vehicle operating in busy, mixed environments the geometric structure of the scene is typically complex, thus we consider $\mathcal{S}$ to represent a pointcloud sampled from the scene, as output from a 3D lidar such as [4], [14].

### B. Formulation of the Optimisation

The appearance of a point $\boldsymbol{q}$ in $\mathcal{S}$ as viewed from $A$ is the value of the image (colour) at the coordinates in the image-plane to which $\boldsymbol{q}$ reprojects and is a function of the intrinsic camera calibration and the position of the camera. Using the pinhole camera model from [15] and denoting the camera parameters by $\boldsymbol{\kappa}$ we can define the reprojection operator $\mathcal{P}$, which maps a point $\boldsymbol{q} \in \mathbb{R}^3$ defined in the same reference coordinate system $R$, as the camera to its image $\boldsymbol{x} \in \mathbb{R}^2$ in the image-plane of the camera. We denote the value of the image (colour) at $\boldsymbol{x}$ by $\mathcal{I}(\boldsymbol{x}) \in \mathbb{R}^k$ for an image with $k$ channels ($k = 3$ for RGB).

$$\boldsymbol{x}_A \equiv \mathcal{P}(\boldsymbol{q}, G_{AR}, \boldsymbol{\kappa}) \tag{1}$$

We can now write our problem shown graphically in Figs. 1 and 3, as that of computing estimates $\widehat{G}_{AR}$ & $\widehat{G}_{BA}$, of $G_{AR}$ and $G_{BA}$ respectively, by minimising an objective function $f : \mathbb{R}^{2\times(|\mathcal{S}_{AB}|\times k)} \mapsto \mathbb{R}^1$ given in eq. (2) (for two $k$-channel images) which measures the discrepancy between the visual appearance of the subset of points $\mathcal{S}_{AB} \subseteq \mathcal{S}$ that reproject into both $\mathcal{I}_A$ and $\mathcal{I}_B$.

$$\begin{aligned} & f\Bigg( \overbrace{\mathcal{I}_A(\mathcal{P}(\boldsymbol{q}, \widehat{G}_{AR}, \boldsymbol{\kappa}))}^{\text{Appearance of } \mathcal{S}_A \text{ from A}}, \overbrace{\mathcal{I}_B(\mathcal{P}(\boldsymbol{q}, \widehat{G}_{BA}\widehat{G}_{AR}, \boldsymbol{\kappa}))}^{\text{Appearance of } \mathcal{S}_B \text{ from B}} \Bigg| \\ & \overbrace{\boldsymbol{q} \in \mathcal{S}_{AB} \equiv \mathcal{S}_A \cap \mathcal{S}_B}^{\text{Scene common to A \& B}} \Bigg) : \mathbb{R}^{2\times(|\mathcal{S}_{AB}|\times k)} \mapsto \mathbb{R}^1 \\ & \equiv f\Big( \mathcal{I}_A(\boldsymbol{x}_A),\ \mathcal{I}_B(\boldsymbol{x}_B) \ \Big|\ \boldsymbol{q} \in \mathcal{S}_{AB} \equiv \mathcal{S}_A \cap \mathcal{S}_B \Big) \end{aligned} \tag{2}$$

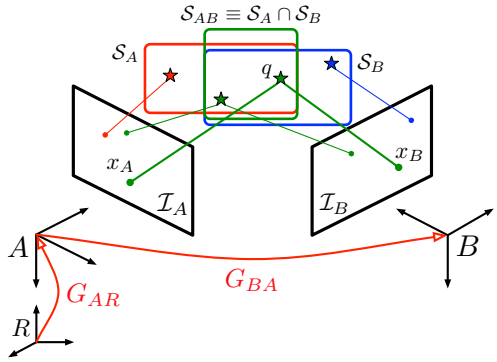



Fig. 3. Two-view reprojection, the camera centers are denoted by the coordinate frames $\{A, B\}$, the images by $\{\mathcal{I}_A, \mathcal{I}_B\}$, the subset of the scene $\mathcal{S}$ visible in each image independently by $\{\mathcal{S}_A, \mathcal{S}_B\}$ and the subset of the scene visible in both images by $\mathcal{S}_{AB}$, which is the only component to contribute to the objective function given by eq. (2).

Two important distinctions between this formulation and current feature-based localisation techniques are the issues of feature selection and correspondance. In feature-based formulations such as [16], [17], the features and their correspondances have to be explicitly identified in each new frame with significant care taken to ensure correct data association. In our formulation the point correspondances between frames are implicitly known through $\mathcal{S}_{AB}$. The issue of point selection is that of selecting the visible subset of points for a given camera position, a typically inexpensive operation that can be aided significantly by offline preprocessing of the pointcloud, which can also be used to improve the suitability of the points in $\mathcal{S}$ for localisation.

From the proposed minimisation given in eq. (3), it is clear that $\mathcal{I}_A(\boldsymbol{x}_A)$ & $\mathcal{I}_B(\boldsymbol{x}_B)$ are dependent upon both $\widehat{G}_{AR}$ and $\widehat{G}_{BA}$. Therefore, so is $\mathcal{S}_{AB} \equiv \mathcal{S}_{AB}|_{\widehat{G}_{AR},\widehat{G}_{BA}}$, for brevity we drop the $|_{\widehat{G}_{AR},\widehat{G}_{BA}}$ nomenclature from $\mathcal{S}_{AB}$, however to be clear $\mathcal{S}_{AB}$ is dependent upon the current candidates for $\widehat{G}_{AR}$ & $\widehat{G}_{BA}$.

$$\{\widehat{G}_{AR}, \widehat{G}_{BA}\} : \min_{\{\widehat{G}_{AR},\widehat{G}_{BA}\}} f\big(\mathcal{I}_A(\boldsymbol{x}_A),\ \mathcal{I}_B(\boldsymbol{x}_B) \mid \boldsymbol{q} \in \mathcal{S}_{AB}\big) \tag{3}$$

As in [16], [5], in the optimisation we estimate $\Delta$-transforms to be applied to initial estimates of the transforms, thus allowing for the incorporation of a motion model, or similar seed to the algorithm.

Numerous potential choices exist for the objective function given in eq. (2). We desire a function that is robust to noise, with a wide basin of convergence to a clear, ideally unique, minimum and is thus well suited to solution by numerical optimisation. In the context of image registration for related problems, [5], [18], [6] use SSD (sum of squared differences), in the case of [5], [18] with efficient second-order minimisation and in [6], with a robust Huber kernel. Whilst [19] uses the $L^1$-norm and [11], [12], [20] maximise the Mutual Information.

We choose instead to use the Normalised Information Distance (NID) [21], [22] as its basis as a Shannon information [23] measure allows a clear, intuitive understanding and it has similar robustness properties to the Mutual Information by virtue of its dependence only upon distributions and not samples. Crucially, unlike the Mutual Information, it is also a true metric.

### C. Normalised Information Distance

The normalised information distance $NID(X, Y)$ for two discrete random variables $\{X, Y\}$ is a Shannon information measure that represents the similarity between the distributions of $X$ and $Y$. Formally, $NID(X, Y)$ is given by eq. (7) where $H(X)$, $H(X, Y)$ and $I(X; Y)$ denote the entropy, joint entropy and mutual information respectively given by eqs. (4) to (6).

$$H(X) \equiv -\sum_{x \in \mathcal{X}} p_x \log(p_x) \tag{4}$$

$$H(X, Y) \equiv -\sum_{x \in \mathcal{X}, y \in \mathcal{Y}} p_{xy} \log(p_{xy}) \tag{5}$$

$$I(X; Y) \equiv H(X) + H(Y) - H(X, Y) \tag{6}$$

The $NID$ is a true metric [22], and is thus non-negative, symmetric and satisfies the triangle inequality: $NID(X, Y) + NID(Y, Z) \geq NID(X, Z)$ and $NID(X, Y) = 0 \iff X = Y$. It is also bounded in both directions: $0 \leq NID(X, Y) \leq 1$, with smaller values indicating greater similarity.

$$NID(X, Y) \equiv \frac{H(X \mid Y) + H(Y \mid X)}{H(X, Y)} \tag{7}$$

$$= \frac{H(X, Y) - I(X; Y)}{H(X, Y)} \tag{8}$$

### D. Application of the Normalised Information Distance

Consider the problem formulation from section III-A, by modeling the appearance of the points in the scene from each view point $\mathcal{I}(\boldsymbol{x})$ as samples from discrete random variables, we can substitute $NID(X, Y)$ as the objective function in eq. (3) to obtain eq. (9). The minimum of which maximises the similarity of the information about the appearance of the scene $\mathcal{S}_{AB}$ provided by $\mathcal{I}_A$ & $\mathcal{I}_B$.

$$\{\widehat{G}_{AR}, \widehat{G}_{BA}\} : \min_{\{\widehat{G}_{AR},\widehat{G}_{BA}\}} NID(\mathcal{I}_A(\boldsymbol{x}_A),\ \mathcal{I}_B(\boldsymbol{x}_B) \mid \boldsymbol{q} \in \mathcal{S}_{AB}) \tag{9}$$

*1) Incorporation of Prior Appearance Information:* A key benefit arising from $NID(X, Y)$ being a true metric, is that if an appearance prior is available for the scene (which we assume will typically be the case) we can meaningfully incorporate it. This is achieved by modifying the objective function to eq. (10), where $\mathcal{I}_{\mathcal{S}}(\boldsymbol{x}_{\mathcal{S}})$ is the prior appearance of a point.

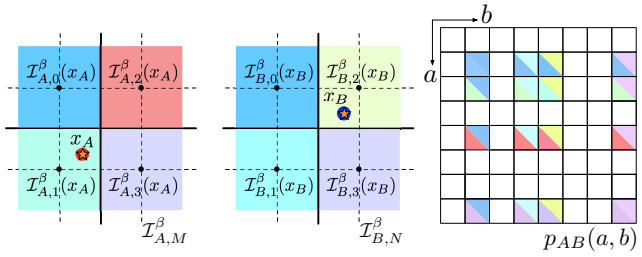



Fig. 4. B-Spline interpolation $p_{AB}$ histogram updates (eq. (11)), note: schematic shown is for 1-degree, not cubic, splines for ease of illustration. $\mathcal{I}^{\beta}_{A,M}$ denotes the $2\times2$ pixel region of local-support (in which the B-spline coefficients $\beta_m(\boldsymbol{x}_A)$ are non-zero), for $\boldsymbol{x}_A$ in the B-spline interpolated image $\mathcal{I}^{\beta}_{A}$ (derived from $\mathcal{I}_A$). Respectively, $\mathcal{I}^{\beta}_{B,N}$ denotes the same concept but for $\mathcal{I}_B$ & $\boldsymbol{x}_B$. For the reprojections of a single 3D point $q$ into both images: $\{\boldsymbol{x}_A, \boldsymbol{x}_B\}$, the bins updated in the histogram approximation of the joint distribution $p_{AB}$ are *not* (necessarily) adjacent. Rather the bins updated are those which contain the values of the pixels in the region of local support for $\{\boldsymbol{x}_A, \boldsymbol{x}_B\}$ in the interpolated images. The contributions to each bin are given by the product of the B-spline coefficients: $\beta_m(\boldsymbol{x}_A)\beta_n(\boldsymbol{x}_B)$ as shown in eq. (11).

$$f \equiv NID(\mathcal{I}_A(\boldsymbol{x}_A),\, \mathcal{I}_B(\boldsymbol{x}_B) \mid \boldsymbol{q} \in \mathcal{S}_{AB}) + NID(\mathcal{I}_A(\boldsymbol{x}_A),\, \mathcal{I}_{\mathcal{S}}(\boldsymbol{x}_{\mathcal{S}}) \mid \boldsymbol{q} \in \mathcal{S}_{AB}) + NID(\mathcal{I}_B(\boldsymbol{x}_B),\, \mathcal{I}_{\mathcal{S}}(\boldsymbol{x}_{\mathcal{S}}) \mid \boldsymbol{q} \in \mathcal{S}_{AB}) \tag{10}$$

Whilst the use of $\mathcal{I}_{\mathcal{S}}(\boldsymbol{x}_{\mathcal{S}})$ might be assumed to be problematic in the event that the camera and/or conditions under which it was captured differ from the current ones. We have found that the $NID$ exhibits similar desirable robustness to illumination changes and occlusion to that obtained by mutual information-based image registration techniques [8], [13]. Specifically, we have found that whilst the use of $\mathcal{I}_{\mathcal{S}}(\boldsymbol{x}_{\mathcal{S}})$ has little impact on the structure of the cost function for $\widehat{G}_{BA}$, it significantly improves the structure for $\widehat{G}_{AR}$, particularly when the inter-image spacing is small and the environment exhibits perceptual aliasing.

*2) Computation of the Joint Appearance Distribution:* There are various methods that could be used to approximate the (discrete) joint appearance distribution: $p_{AB}(a,b) \equiv \Pr(\mathcal{I}_A(\boldsymbol{x}_A) = a \cap \mathcal{I}_B(\boldsymbol{x}_B) = b)$. As we wish to utilise gradient-based optimisation methods, we require a form for $p_{AB}$ with analytical derivatives which are fast to compute. Following from [24], [25], we use cubic B-Spline [26] interpolation to update multiple bins in the joint-histogram for each $\boldsymbol{x}$, with fractional weights corresponding to the product of the B-Spline coefficients; represented graphically in Fig. 4.

$$p_{AB}(a,b) = \frac{1}{|\mathcal{S}_{AB}|} \sum_{q \in \mathcal{S}_{AB}} \sum_{m \in M} \sum_{n \in N} \beta_m(\boldsymbol{x}_A)\beta_n(\boldsymbol{x}_B)\, \delta\Big(a - \Big\lfloor \frac{\mathcal{I}^{\beta}_{A,m}}{\tau_A} \Big\rfloor\Big)\delta\Big(b - \Big\lfloor \frac{\mathcal{I}^{\beta}_{B,n}}{\tau_B} \Big\rfloor\Big) \tag{11}$$

Formally, we write the histogram approximation of $p_{AB}$ as a summation over all points given in eq. (11). Where $\beta_m(\boldsymbol{x}_A)$ denotes the B-spline coefficient for the $m-$th pixel in the region of local-support for $\boldsymbol{x}_A$ (in which the B-spline coefficients are non-zero) in the interpolated image $\mathcal{I}^{\beta}_{A}$. By definition, the cardinality of the region of local-support ($M$) is defined by the order of the B-spline, for the cubic B-splines used here $|M| = 16$. Respectively, the other terms denote the same concepts, but for the image captured at $B$.

$$\sum_{m \in M} \beta_m(\boldsymbol{u}) = 1, \qquad \{\beta_m(\boldsymbol{u}) = 0 \quad \forall\, m \notin M\} \tag{12}$$

In the context of the standard form for a B-spline surface consisting of a double summation over the control points in each parametric direction, $\beta_m(\boldsymbol{x}_A)$ represents the product of the two univariate B-spline basis functions and hence satisfies the partition of unity property eq. (12), thus eq. (11) requires no additional normalisation.

The delta-functions in eq. (11) determine which bin is updated in the joint histogram, $\mathcal{I}^{\beta}_{A,m}$ denotes the value of the $m-$th pixel in the region of local-support for $\boldsymbol{x}_A$ in the interpolated image and $\tau_A$ denotes the bin-size. In the results presented here, we have used evenly spaced bins across the range of values for each channel $(0-255)$, with the number of bins $\eta_A = \eta_B = \eta = 30$, however we have found that the formulation presented here is typically very robust to changes in $\eta$.

*3) Computation of Analytical Derivatives:* In a manner similar to that of [24], [25] for the mutual information, we can compute analytical derivatives for eq. (9). Differentiating the objective function with respect to a transform $G$ yields eq. (13), where we have abbreviated the nomenclature for brevity. The corresponding derivatives for the joint-information $H_{AB}$ and mutual information $I_{AB}$ are given in eqs. (14) and (15). We have specialised eq. (15) to be with respect to $G_{BA}$ as its form is not equivalent (excepting the $\partial p_{AB}$ term) under $\partial G_{AR}$, as in that case $I_{AB}$ depends on $p_A$ which is not constant.

$$\frac{\partial NID_{AB}}{\partial G} = \frac{H_{AB}\big(\frac{\partial H_{AB}}{\partial G} - \frac{\partial I_{AB}}{\partial G}\big) - \frac{\partial H_{AB}}{\partial G}\big(H_{AB} - I_{AB}\big)}{H^2_{AB}} \tag{13}$$

$$\frac{\partial H_{AB}}{\partial G} = -\sum_{a,b} \frac{\partial p_{AB}}{\partial G}\Big(1 + \log(p_{AB})\Big) \tag{14}$$

$$\frac{\partial I_{AB}}{\partial G_{BA}} = \sum_{a,b} \frac{\partial p_{AB}}{\partial G_{BA}}\Big(1 + \log(\frac{p_{AB}}{p_B})\Big) \tag{15}$$

Considering eq. (11) and noting that $\beta(\boldsymbol{x})$ depends only on $\boldsymbol{x}$, *not* on $\mathcal{I}^{\beta}$ and that $\boldsymbol{x}_A$ is not dependent upon $G_{BA}$, $\frac{\partial p_{AB}}{\partial G_{BA}}$ is another histogram summation involving eq. (16) (and $\beta_m(\boldsymbol{x}_A)$).

$$\frac{\partial \beta_n(\mathcal{P}(\boldsymbol{q}_R, G_{BA}G_{AR}))}{\partial G_{BA}} = \frac{\partial \beta_n}{\partial \boldsymbol{x}_B}\frac{\partial \boldsymbol{x}_B}{\partial \boldsymbol{q}_B}\frac{\partial \boldsymbol{q}_B}{G_{BA}} \tag{16}$$

Where $\frac{\partial \beta_n}{\partial \boldsymbol{x}_B}$ is obtained from the standard B-spline basis function derivative equations [26] and $\frac{\partial \boldsymbol{x}_B}{\partial \boldsymbol{q}_B}$ is dependent

upon the camera model, for which in the work presented here we have used the formulation presented in [27].

The derivatives of eq. (10) with respect to $G_{BA}$ and eqs. (9) and (10) with respect to $G_{AR}$ can also be similarity computed, via modest changes to $\partial I_{AB}$ and $\frac{\partial \boldsymbol{q}}{\partial G}$.

*4) Extension to Multiple Cameras:* The extension of the method to a vehicle with multiple cameras follows naturally by exploiting the metric property of $NID$. Assuming the images are temporally aligned, the $NID$ is computed between consecutive frames for each camera independently, then summed over all cameras to produce the total cost.

*5) Expanding Convergence Basin:* As noted in [8], [13] for the alignment of images using the mutual information, blurring the images prior to alignment typically results in a broadening of the convergence basin. Empirically we have found that this also applies to the $NID$.

## IV. Results

To evaluate the performance of the proposed system, a 3D lidar pointcloud $\mathcal{S}$, consisting of 2.6 million points was constructed for the loop with a circumference of approximately 690m shown in Fig. 5 using a SICK LMS-151 lidar mounted on the roof of a vehicle, approximately 2m off the ground and declined by $8.5^o$. The vehicle was also equipped with a Point Grey Bumblebee2 monochrome stereo camera, from which only the images from the left camera were used; and an OxTS 3042 6-DoF INS used for ground-truth. The images from the left camera of the stereo pair for this loop (after rectification) form the test image set, consisting of 271 512x384 images sampled at 2Hz at a speed of 10mph resulting in a mean image separation of 2.5m. The camera data from a second loop was used to build the appearance prior for $\mathcal{S}$, taken to be the average intensity for each point across all images captured within 60m in which the point was visible.

We adopt the NASA coordinate system as used in [16] with X-forward, Y-right and Z-down in the robot's coordinate frame. An offset UTM coordinate system with X-Northing, Y-Easting, Z-Down is used as the reference coordinate frame $R$, in which $\mathcal{S}$ is defined and the robot's pose is to be estimated.

Fig. 6 shows iso-surface plots of the cost function components of eq. (10) for $G_{BA}$ & $G_{AR}$ separately, for $\Delta$s about their ground-truth values given in table I for the two sequential images shown in Figs. 6a and 6c. With the transform for which the cost is not shown in each figure held at its ground-truth value in each case. The location of the first image in $\mathcal{S}$ is highlighted in Fig. 5.

Considering the iso-surfaces for $\Delta G_{BA}$ in Figs. 6e and 6f, the cost function is the local inter-frame cost component of eq. (10): $NID(\mathcal{I}_A, \mathcal{I}_B)$ and appears very well behaved and well suited to optimisation: smooth, with clear minima at the ground-truth values ($\Delta = 0$), shown by the sample lines evaluated along each $\Delta$ axis. Additionally, the analytical gradients evaluated along the sample lines are also well behaved and accurately represent the gradient of the cost.

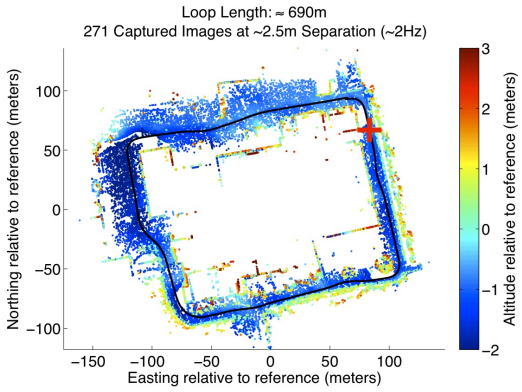


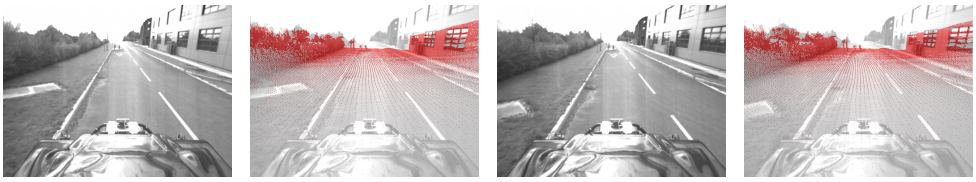

Fig. 5. Overhead view of $\mathcal{S}$ used for experimental results, consisting of 2.6 million points (subsampled for display) defined in an offset UTM coordinate system, with the loop trajectory shown in black and the location of the images used for Fig. 6 denoted by a red cross.

The elongation of the iso-surfaces in Fig. 6e along the $\Delta$X axis corresponding with the wide, shallow peak in the sample line along its axis, is a result of the camera and scene geometry. From Figs. 6b and 6d it is clear that neither the camera view, nor the position of the reprojected points shifts significantly between the images, despite $G_{BA}$ consisting of a translation of 2.2m in the X direction, thus a shallow peak in this axis is intuitive. The sharper peaks in Fig. 6f are similarly intuitive given the greater affect on the reprojected positions of points in the far-field of angular changes given the position of the camera on the vehicle.

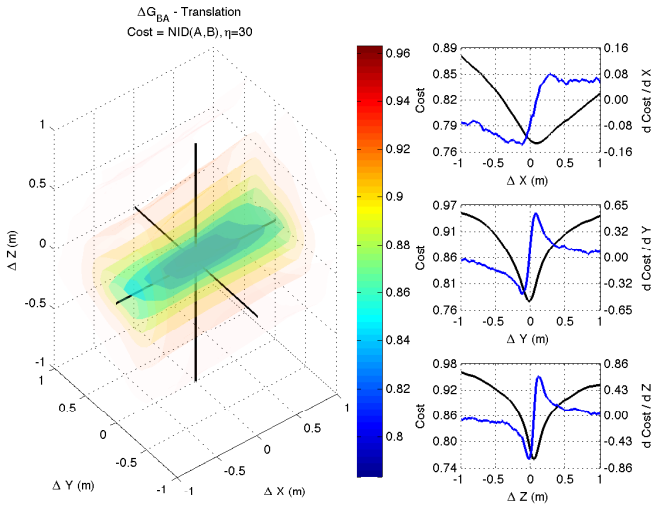
The results for $\Delta G_{AR}$ are shown in Figs. 6g and 6h, where the cost shown is the local-to-prior component of eq. (10): $NID(\mathcal{I}_A, \mathcal{I}_\mathcal{S}) + NID(\mathcal{I}_B, \mathcal{I}_\mathcal{S})$, with $G_{BA}$ held at its ground truth value. This function and its analytical derivatives also appear relatively well behaved, with clear minimum in the cost at the ground-truth values ($\Delta = 0$). The weak constraint angled at approximately $10^o$ off the X-axis in the XY plane in Fig. 6g is the mapping of the weak constraint in X in Fig. 6e from the local robot coordinate system (X forward) into the reference coordinate system of $G_{AR}$ (X-North). As at the points the images were captured, the vehicle was travelling within $10^o$ of due south in the reference frame, as shown in Fig. 5.

These results are indicative of the robustness of $NID$ to changing conditions, as the simple (averaged) appearance prior is not very accurate or even very consistent, due to illumination and auto-exposure changes between the widely spaced images.

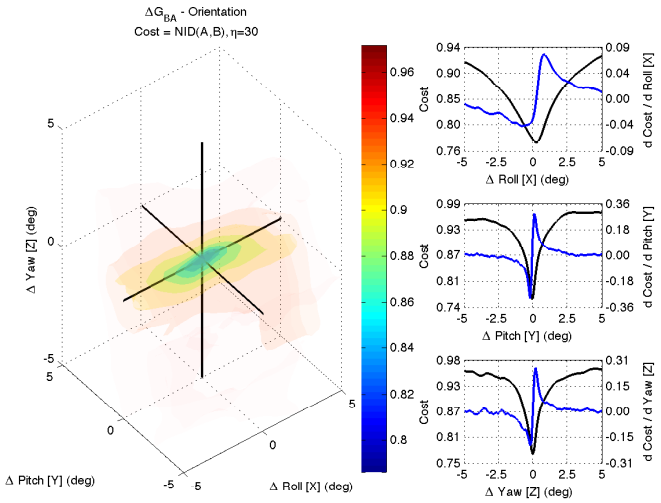
In order to investigate the convergence of $\widehat{G}_{BA}$, independently for each pair of consecutive images in the test set $G_{AR}$ was held at its ground-truth value and $\widehat{G}_{BA}$ was initialised to its ground-truth value with samples drawn from independent uniform additive noise with extents $\pm 0.5m$ and $\pm 2.5^o$ added to the translation and orientation components respectively. From this noisy initial position a quasi-Newton optimiser using the analytical derivatives detailed in section III-D.3 was used to find the minima, stopping when $\Delta\text{Cost} \leq 10^{-6}$. Two sequential optimisation cycles were performed for each



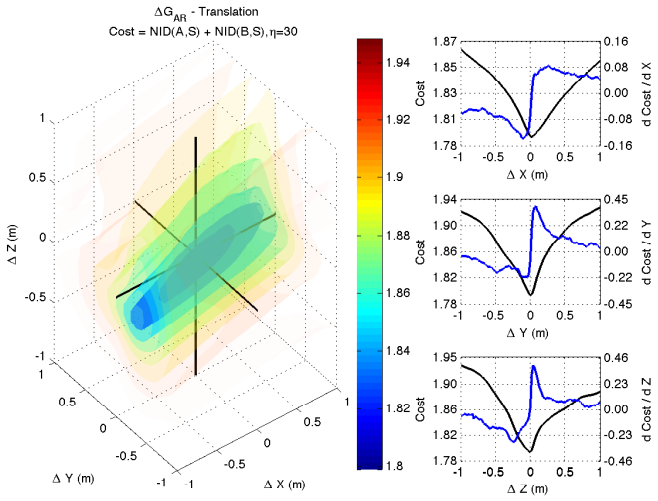
(a) (b) (c) (d)



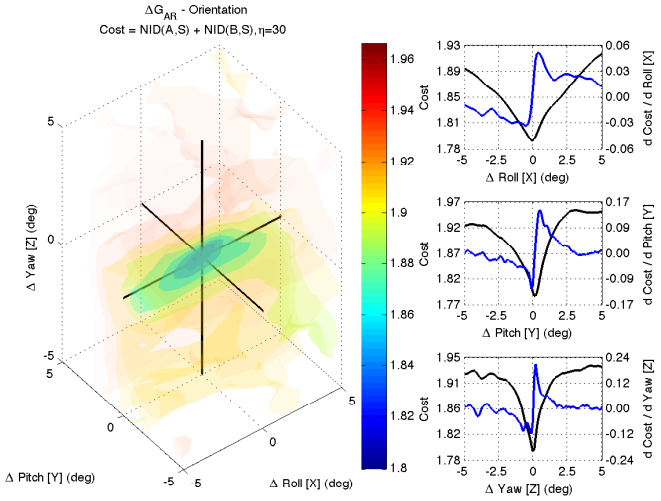

(e) $\Delta G_{BA}$ X-Y-Z Cost Volume




(f) $\Delta G_{BA}$ Roll-Pitch-Yaw Cost Volume




(g) $\Delta G_{AR}$ X-Y-Z Cost Volume




(h) $\Delta G_{AR}$ Roll-Pitch-Yaw Cost Volume

Fig. 6. Typical iso-surface cost structures for $\Delta G_{AR}$ & $\Delta G_{BA}$, sampled around ground-truth values ($\Delta = 0$) with the transform not shown held at its ground-truth value. Approximately 30,000 prior lidar points are common to both sample images ($\mathcal{S}_{AB}$): 6b and 6d and were used in the computation of the costs. The subplots on the right of each figure show the evolution of the cost (black) and its analytical derivative (blue) along the corresponding black sample lines aligned with each $\Delta$-axis in the iso-surface plot. The iso-surfaces have been inverse alpha-blended (higher costs are more transparent) for ease of illustration.

| Transform | X (m) | Y (m) | Z (m) | Roll (deg.) | Pitch (deg.) | Yaw (deg.) |
|---|---|---|---|---|---|---|
| $G_{AR}$ | 49.3473 | 94.8116 | -0.6958 | -2.2266 | -3.0308 | -168.7081 |
| $G_{BA}$ | -2.2108 | 0.0559 | 0.1071 | -0.0647 | 0.0990 | 0.0308 |

TABLE I

GROUND-TRUTH TRANSFORM COMPONENTS FOR RESULTS SHOWN IN FIG. 6.

image pair, the first using blurred images (gaussian filter, 15 pixels square with $\sigma = 5$). The second cycle, which took as its input the output of the first, used the raw (unmodified) images. This was done to improve the basin of convergence as noted in section III-D.5 by approximating an image-pyramid technique. As the aim was to investigate $G_{BA}$, the scene prior was not used.

Figs. 7a to 7c show the distributions over all sequential test image pairs, of the initial additive noise and residual errors in $\widehat{G}_{BA}$ at the end of each cycle of the optimisation respectively. The results in Fig. 7c broadly reflect the width of the peaks at the minima in Figs. 6e and 6f, with greater confinement in the orientation components than in the translation components. Contributors to the width of the error peaks include: residual error in the extrinsic camera calibration used to estimate ground-truth, the low resolution of the camera and some particularly challenging areas where the camera views are non-informative (looking into a hedge). However, the error distributions, particularly in orientation provide strong evidence of the validity of the formulation and to the typicality of the results presented in Figs. 6e and 6f.

As our current implementation is off-line and not capable of real-time performance, to demonstrate the speed of convergence of the optimisation Figs. 7d and 7e show respectively, a histogram over all image-pair samples of the 'cost-to-come' at each iteration and the percentage of samples for which the optimiser had not terminated after $k$-iterations. The key point to note from Fig. 7d is that the 'cost-to-come' converges to zero rapidly after $\approx 10$ iterations, with the remainder of the time spent performing fine-detail refinement.

## V. Conclusions and Future Work

In this paper we have presented a new method for localising a robot equipped with a monocular camera in a previously gathered 3D lidar pointcloud survey prior by harmonising both the locally viewed appearance of the workspace between consecutive frames and the local appearance of the workspace relative to an appearance prior. Our method uses the Normalised Information Distance ($NID$), a true Shannon information metric, as the cost function to compare distributions over appearance for common subsets of $\mathcal{S}$ and is suitable for real-time, parallel computation. We have presented results demonstrating the desirable behaviour of the $NID$ as a cost function for our problem using data gathered from a road vehicle driving in a light-urban environment.

The primary foci for our future work are twofold. Firstly, to accelerate our current implementation to achieve real-time, online performance. In part by using the work of [28] for GPU accelerated B-spline interpolation, which is a significant performance bottleneck in our current implementation. Secondly, we hope to exploit multiple cameras on the robot to be localised with different view-points to improve the gradient of the cost function for the translation components of the transforms. As previously discussed, the confinement of the $NID$ cost function in our formulation is a function of both the scene and the relative geometry of the camera(s), thus an additional focus of this work is to investigate pre-processing of the raw pointcloud to improve its suitability and evaluate its sufficiency for navigation.

We feel this work demonstrates the feasibility of our approach and supports our expectation that the localisation systems of practical future autonomous vehicles will benefit from making extensive use of prior 3D information. Furthermore, that by doing so they could reduce their sensing requirements and costs whilst maintaining the accuracy and robustness required.

## VI. Acknowledgements

This work was generously supported by EPSRC through a DTA Studentship supporting Alex Stewart and EPSRC Leadership Fellowship EP/I005021/1 supporting Paul Newman.

## References

[1] D. Anguelov, C. Dulong, D. Filip, C. Frueh, S. a. Lafon, R. Lyon, A. Ogale, L. Vincent, and J. Weaver, "Google street view: Capturing the world at street level," *Computer*, vol. 43, no. 6, pp. 32 –38, 2010.

[2] C. Urmson *et al.*, "Autonomous driving in urban environments: Boss and the Urban Challenge," *Journal of Field Robotics*, vol. 25, no. 8, pp. 425–466, 2008.

[3] J. Leonard *et al.*, "A Perception-Driven Autonomous Urban Vehicle," *Journal of Field Robotics*, vol. 25, no. 10, pp. 727–774, 2008.

[4] M. Sheehan, A. Harrison, and P. Newman, "Self-calibration for a 3d laser," *The International Journal of Robotics Research*, 2011.

[5] G. Silveira, E. Malis, and P. Rives, "An efficient direct approach to visual SLAM," *Robotics, IEEE Transactions on*, vol. 24, no. 5, pp. 969–979, 2008.

[6] A. Comport, E. Malis, and P. Rives, "Real-time Quadrifocal Visual Odometry," *The International Journal of Robotics Research*, vol. 29, no. 2-3, p. 245, 2010.

[7] R. Newcombe, S. Lovegrove, and A. Davison, "DTAM: Dense tracking and mapping in real-time," in *Computer Vision, 2011. ICCV 2011. IEEE International Conference on*, 2011.

[8] J. P. W. Pluim, J. B. A. Maintz, and M. A. Viergever, "Mutual-information-based registration of medical images: a survey," *Medical Imaging, IEEE Transactions on*, vol. 22, no. 8, pp. 986–1004, 2003.

[9] F. Maes, A. Collignon, D. Vandermeulen, G. Marchal, and P. Suetens, "Multimodality image registration by maximization of mutual information," *Medical Imaging, IEEE Transactions on*, vol. 16, no. 2, pp. 187–198, 1997.

[10] F. Maes, D. Vandermeulen, and P. Suetens, "Comparative evaluation of multiresolution optimization strategies for multimodality image registration by maximization of mutual information," *Medical Image Analysis*, vol. 3, no. 4, pp. 373 – 386, 1999.

[11] G. Panin and A. Knoll, "Mutual information-based 3D object tracking," *International Journal of Computer Vision*, vol. 78, no. 1, pp. 107–118, 2008.

[12] A. Dame and E. Marchand, "Mutual Information-Based Visual Servoing," *Robotics, IEEE Transactions on*, vol. 27, no. 5, pp. 958–969, 2011.

[13] A. Dame and E. Marchand, "A new information theoretic approach for appearance-based navigation of non-holonomic vehicle," in *Robotics and Automation (ICRA), 2011 IEEE International Conference on*, pp. 2459–2464, 2011.

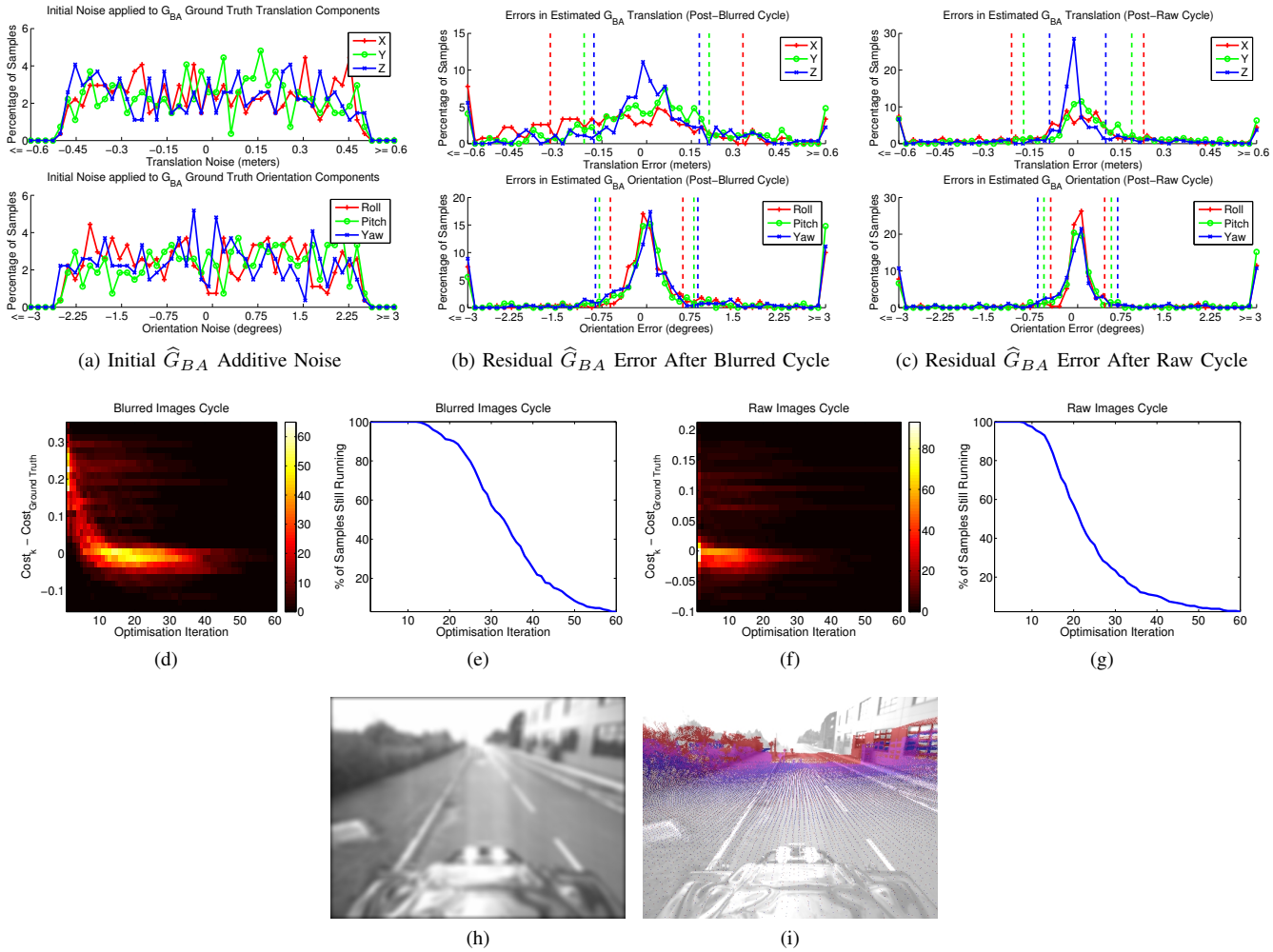
[14] Velodyne Lidar Inc., *Velodyne HDL-64E: A High Definition Lidar Sensor for 3-D Applications*, 2007.

[15] R. I. Hartley and A. Zisserman, *Multiple View Geometry in Computer Vision*. Cambridge University Press, ISBN: 0521623049, 2000.

[16] G. Sibley, C. Mei, I. Reid, and P. Newman, "Vast-scale Outdoor Navigation Using Adaptive Relative Bundle Adjustment," *The International Journal of Robotics Research*, vol. 29, no. 8, p. 958, 2010.

[17] K. Konolige, J. Bowman, J. Chen, P. Mihelich, M. Calonder, V. Lepetit, and P. Fua, "View-based maps," *The International Journal of Robotics Research*, vol. 29, no. 8, p. 941, 2010.

[18] G. Silveira and E. Malis, "Unified Direct Visual Tracking of Rigid and Deformable Surfaces Under Generic Illumination Changes in Grayscale and Color Images," *International Journal of Computer Vision*, vol. 89, pp. 84–105, Aug. 2010.




(a) Initial $\widehat{G}_{BA}$ Additive Noise (b) Residual $\widehat{G}_{BA}$ Error After Blurred Cycle (c) Residual $\widehat{G}_{BA}$ Error After Raw Cycle

(d) (e) (f) (g)

(h) (i)

Fig. 7. Results of independent optimisations ($\eta = 30$) for $G_{BA}$ (with $G_{AR}$ held at its ground-truth value) over all sequential pairs of images in the test-set. Showing the distributions of additive noise added to ground-truth to form initialization values of $\widehat{G}_{BA}$: 7a and the residual errors in $\widehat{G}_{BA}$ after termination: 7b and 7c. The vertical dashed lines in 7b and 7c denote the 1-$\sigma$ bounds. The speed of convergence is shown in 7d to 7g, which show for the blurred & raw (unmodified) image cycles, a histogram over all image pair samples of the 'cost-to-come' at each iteration and the percentage of optimisations still running after $k$-iterations respectively. Finally, 7h shows a sample image after blurring (gaussian filter, 15 pixels square with $\sigma = 5$) and 7i shows one of the test-images with point reprojections using the ground-truth $G_{BA}$ (red) and a noisy initialisation [before optimisation] (blue) drawn from 7a. In the case shown, the additive noise on $G_{BA}$ is $[-0.225, -0.065, 0.033]$ meters $(X, Y, Z)$ and $[-1.383, -2.248, 2.135]$ degrees (Roll, Pitch, Yaw). The point reprojections after the optimisation are not shown as they are indistinguishable from those at ground-truth


[19] T. Pock, M. Urschler, C. Zach, R. Beichel, and H. Bischof, "A duality based algorithm for TV-$L_1$-optical-flow image registration," in *Medical Image Computing and Computer-Assisted Intervention MICCAI 2007*, vol. 4792 of *Lecture Notes in Computer Science*, pp. 511–518, Springer Berlin / Heidelberg, 2007.

[20] P. Viola and W. I. Wells, "Alignment by maximization of mutual information," in *Computer Vision, 1995. Proceedings., Fifth International Conference on*, pp. 16–23, 1995.

[21] M. Li, X. Chen, X. Li, B. Ma, and P. Vitanyi, "The similarity metric," *Information Theory, IEEE Transactions on*, vol. 50, pp. 3250 – 3264, dec. 2004.

[22] A. Kraskov, H. Stögbauer, R. G. Andrzejak, and P. Grassberger, "Hierarchical Clustering Based on Mutual Information," *eprint arXiv:q-bio/0311039*, Nov. 2003.

[23] C. E. Shannon, "A Mathematical Theory of Communication," *The Bell System Technical Journal*, vol. 27, pp. 379–423, 623–656, 1948.

[24] N. Dowson and R. Bowden, "A unifying framework for mutual information methods for use in non-linear optimisation," in *Computer Vision - Eccv 2006 , Pt 1, Proceedings*, (Univ Surrey, Ctr Vis Speed & Signal Proc, Guildford GU2 JW, Surrey, England), pp. 365–378, Univ Surrey, Ctr Vis Speed & Signal Proc, Guildford GU2 JW, Surrey, England, 2006.

[25] P. Thevenaz and M. Unser, "Optimization of mutual information for multiresolution image registration," *Image Processing, IEEE Transactions on*, vol. 9, no. 12, pp. 2083–2099, 2000.

[26] L. Piegl and W. Tiller, *The NURBS book*. Springer Verlag, 1997.

[27] G. Klein, *Visual Tracking for Augmented Reality*. PhD thesis, University of Cambridge, 2006.

[28] D. Ruijters and P. Thevenaz, "GPU Prefilter for Accurate Cubic B-spline Interpolation," *The Computer Journal*, vol. 55, pp. 15–20, Dec. 2011.